소재 연구데이터의 구조 및 표준 어휘 사전 구축[1)]

김수현[1], 김정한[2], 김지한[3], 신호선[4], 안재평[5], 오창석[6], 이광렬[7]*, 이동화[8], 이정훈[7], 박선화[4], 장현주[9], 한상수[7], 한승우[10], 허용학[11] (가나다순)

[1]한국재료연구원 금속재료연구본부, [2]국립한밭대학교 신소재공학과, [3]한국과학기술원 생명화학공학부,
[4]한국표준과학연구원 스마트소자팀, [5]한국과학기술연구원 연구자원데이터지원본부
[6]한국재료연구원 재료디지털플랫폼연구본부, [7]한국과학기술연구원 첨단소재연구본부 계산과학연구센터
[8]포항공과대학교 신소재공학과, [9]한국화학연구원 화학데이터기반연구센터,
[10]서울대학교 공과대학 재료공학부 및 신소재공동연구소, [11]국가소재연구데이터센터 한국표준과학연구원

*E-mail: krlee@kist.re.kr

초록

체계적으로 구조화된 소재 연구 데이터는 그 자체로 활용도가 높은 연구개발 자원일 뿐 아니라, 인공지능에 의한 소재 설계 역량을 갖추기 위한 핵심 연구기반이다. 다양한 소스로부터 수집된 소재 연구 데이터를 효율적으로 검색하고 활용하기 위해서는 데이터의 이름 등 키워드를 통일하고 해당 데이터의 타입 그리고 수치 데이터의 단위를 일치시키는 것이 대단히 중요하다. 또한, 머신러닝에 의한 소재 설계를 위해서는 소재의 성능 데이터를 소재의 구조나 공정상의 특이성과 연계하여 수집 관리하여야 한다. 이 상관관계를 기계적으로 학습함으로써 소재 설계의 머신러닝 모델이 만들어질 수 있기 때문이다. 따라서, 데이터 키워드의 표준화와 함께 데이터의 구조가 표준화되어야 소재 연구 데이터의 활용성을 크게 증진시킬 수 있을 것이다. 소재 연구 데이터 표준화 전문위원회에서는 다양한 소재 연구 데이터들을 일관되게 다룰 수 있는 데이터 구조를 구축하여 제안하였다. 응용범위가 다양한 소재 분야의 특성을 고려하여 "소재 시스템" 개념을 채택한 소재 연구 데이터의 표준 구조를 제안하고, 데이터 키워드 어휘들을 정의한 표준 어휘 사전을 구축하여 공개하고 있다.

I. 서론

현대 초연결 사회를 실현한 ICT 기술의 비약적 발전은 소재 분야에서도 연구개발 방법론 자체에 큰 변화를 일으키고 있다. 합성과 분석을 반복하는 실험적 연구와 열역학 및 kinetics에 바탕을 둔 이론적 해석연구가 소재 개발의 고전적인 접근이었다면, ICT 기술의 혁신적 발전은 전자, 원자 수준의 계산과학을 이용한 소재 설계와 데이터에 기반한 소재 정보학 등 새로운 개념의 효율적인 연구개발 수단을 제공하고 있다. 방대한 양의 소재 데이터는 이를 체계적으로 분석하는 것만으로도 소재 탐색의 방향성을 제시할 수 있기 때문에 실험 연구의 효율성을 크게 향상시킬 수 있을 것으로 기대하고 있다. 최근에는 한 발짝 더 나아가 축적

1) 본 논문은 한국연구재단의 지원을 받는 국가 소재 연구데이터 센터, 소재 연구데이터 표준화 전문위원회의 연구 결과이다.

된 R&D 데이터와 첨단의 딥러닝 알고리즘을 접목하는 데이터 기반의 소재 개발 연구가 많은 관심을 끌고 있다.[2] 특히, 연구 및 분석 장비의 현대화를 통해 디지털 형태의 소재 데이터가 급증하면서 그 중요성은 더욱 커지고 있다. 소재 연구는 본질적으로 구조적 특징과 조성적 특징을 제어하여 필요로 하는 기능의 소재를 찾아내는 과정이라고 할 수 있다. 따라서, 딥러닝 기법을 이용하여 소재의 구조적 특징과 조성적 특징 그리고 이를 제어하는 공정 데이터와 해당 소재의 성능 데이터의 상관관계를 규명할 수 있다면, 소재 데이터의 가치가 크게 높아질 것이다. 최근 급증하고 있는 머신러닝에 의한 소재 설계 논문들은 이러한 기대의 실현 가능성을 잘 보여주고 있다.[3]

소재 정보학 혹은 머신러닝에 의한 소재 설계의 필요조건은 무엇보다도 대량의 품질 좋은 소재 데이터를 확보하는 것이다.[4] 이러한 데이터가 존재하지 않으면 아무리 훌륭한 머신러닝 능력으로도 실질적인 가치를 창출할 수 없기 때문이다. 현재 국내의 소재 연구 데이터는 대부분 개별 연구자(팀)이 소유하고 있고, 논문과 보고서의 출판을 통해 데이터가 공개되는 수준에 머물러 있다. 게다가 국가 R&D 사업의 성과물인 연구사업 보고서, 논문 그리고 특허 등을 단순히 국가과학기술정보서비스 (National Science & Technology Information Service; NTIS)에 등록하는 데 그치고 있을 뿐 데이터 기반의 연구에 활용할 수 있는 구조화된 형태로 보관되어 있지 않다. 따라서, 향후 소재 정보학의 발전을 위해서는 흩어져 있는 소재 데이터를 머신러닝 등 소재 정보학에서 손쉽게 활용할 수 있는 형태로 수집하고 관리하는 시스템의 구축이 긴요하다. 또한, 2017년 이후 소재 데이터의 축적과 관리를 위한 선행 연구들이 개별적으로 진행되고 있어 이를 통합하기 위한 노력도 시급한 시점이다.[5]

본 연구에서는 머신러닝에 의한 소재 설계 역량을 강화하기 위해 국가적 규모의 소재 연구개발 데이터의 수집, 관리 및 활용의 표준 체계를 수립하고자 하였다. 여기서 표준화는 소재 데이터 키워드의 어휘를 통일하고 해당 데이터의 타입 그리고 수치 데이터의 단위를 규정하는 것을 의미한다. 데이터는 일반적으로 데이터 키워드와 해당 데이터의 짝으로 구성된다. 예를 들어 {"hardness":58}[6] 인 데이터의 경우 "hardness"가 데이터 키워드이며 58은 이에 해당되는 데이터이다. 이 예시 데이터는 표준화 작업을 통해 정의된 "hardness"라는 키워드를 사용하며, "hardness"에 해당되는 데이터의 타입 (이 예에서는 수치) 과 단위 (이 예에서는 Hv, 비커스 경도값) 에 따라 실제 데이터 "58 Hv"가 입력된 것이다. 표준화된 키워드를 사용하여 수집된 데이터들만이 활용단계에서 체계적으로 검색되고 취득될 수 있기 때문에 키워드의 표준화는 데이터의 활용을 위해 매우 중요한 의미를 갖는다. (이하 어휘는 키워드 어휘를 의미한다.) 최근 소재 데이터 어휘 표준

2) 김동훈, 히로시 미즈세키, 이광렬, "빅데이터와 소재연구", 소재기술백서2017, p.589, 2017.12, 한국재료연구소
3) 한상수, 김동훈, "인공지능 기반 소재설계 연구동향", 재료마당 33, 50 (2020).
4) 여기서 의미하는 품질은 데이터의 신뢰성 뿐 아니라 데이터가 머신러닝에 활용될 수 있는 형태로 구조화되었는지를 의미한다.
5) 오창석, 김수현, 박지원, 이호원, 임창동, 미래소재 연구데이터 플랫폼 구축사업 상세기획 연구 보고서, 2019.4.15. 과학기술정보통신부
6) 본 연구에서 소재 표준 어휘는 모두 영문으로 정의하였다. 데이터 기반 소재 설계의 성공 여부는 가용한 데이터의 규모에 의존하므로, 소재 R&D 빅데이터의 활용 여부가 데이터 기반 소재 연구의 핵심 경쟁력이 될 것이다. 따라서, 대량의 소재 데이터를 활용하기 위한 국제적 협력에 우리나라도 주도적으로 참여하기 위해 영문 표준어휘의 소재 데이터 관리가 필요하다고 판단하였다. 국내 데이터 기반 소재 연구의 확대와 파급을 위해서는 영문 표준어휘를 바탕으로 한 국문 표준어휘 사전을 추가로 구축해야 할 것이다. 또한, 모든 어휘는 혼란을 방지하기 위해 고유명사 외에는 모두 소문자를 사용하였다. 데이터 이름 어휘의 예를 들면 "chemical information", "Czochralski growth method" 등이다.

화를 위한 노력이 전 세계적으로 진행되고 있는데, 그 대표적인 활동으로 Research Data Aliance (RDA) 산하 International Materials Resource Registries Working Group[7]의 노력을 들 수 있다. 이 워킹그룹은 2017년 7월 초기 단계 수준의 "materials registry vocabulary"의 초안을 발표한 바 있다. 그러나, 그 이후의 진전은 매우 더딘 상태이다. 본 연구의 어휘 사전을 바탕으로 기존의 국제 표준화 노력들과 협력하여 소재 데이터 국제표준의 초석을 만드는 데 기여할 수도 있을 것이다.

머신러닝에 의한 소재 설계는 소재의 성능 데이터와 소재의 구조, 공정상의 특이성이 가진 다차원의 상관관계를 학습함으로써 가능해 진다. 소재 데이터의 활용 단계에서 이 상관관계를 체계적으로 다루기 위해서는 잘 정의된 데이터 구조 (데이터 스키마)를 바탕으로 서로 연관된 데이터들이 수집 관리되어야 한다. 따라서, 어휘 자체의 표준화와 함께 소재 데이터의 구조 표준화 역시 데이터 활용을 위해 매우 중요한 요소이다. 본 연구에서는 어휘의 표준화에 선행하여 소재 R&D 데이터의 표준구조를 구축하였다.

일반적으로 소재 데이터는 소재의 활용 영역에 따라 매우 다양한 형태의 조성, 구조, 공정 그리고 특성 정보로 이루어진다. 2012년 미국의 The Minerals, Metals and Materials Society (TMS), American Societry of Metals (ASM)과 미국립표준연구소 (National Institute of Standards & Technology: NIST)가 공동으로 소재 데이터 범주를 음향 데이터부터 열화학 데이터에 이르기까지 30가지로 분류한 바 있다.[8] 그러나, 첨단 소재의 개발과 융합연구가 활성화되면서 이러한 범주는 계속 확장되고 있고 소재의 특이성과 물성을 바라보는 관점 또한 매우 다양해 지고 있다. 따라서, 표준화된 소재 데이터 구조는 광범위한 영역의 소재 데이터들을 포괄적으로 다룰 수 있어야 할 뿐 아니라 확장성도 가져야 한다. 이러한 소재 분야의 특성을 반영하여 2020년 발족한 국가소재연구데이터센터는 소재를 크게 에너지 환경 소재, 스마트 IT 소재 그리고 구조 안전 소재 등 3개의 영역으로 구분하고 각 영역 내에서 특화된 소재군의 전문가들이 소재연구데이터 구축에 참여하고 있다. 본 연구에서는 현재의 국가소재연구데이터센터가 다룰 소재 데이터를 포괄적으로 수용할 수 있는 데이터 구조를 제안하고 데이터 수집과 활용을 위한 키워드와 해당 데이터 형태를 정의하였다. 그러나, 소재 연구데이터의 범위가 현재의 3개 영역으로 제한되지 않고 계속 확장될 것이므로 향후 확장성을 고려한 데이터 구조를 설계하고자 하였다. 그 결과, 우리는 "소재 시스템"의 관점에서 소재 데이터 구조를 제안하고, 이를 바탕으로 표준소재어휘사전을 구축할 수 있었다.

II. 소재 시스템의 개념과 소재 데이터 구조

본 연구에서는 다양한 소재 데이터들을 포괄적으로 수용하기 위해 "소재 시스템"의 개념을 도입하여 소재 데이터를 구조화하였다. 소재 시스템이란 특정 소재의 성능을 평가할 수 있도록 한 개 이상의 서로 다른 소재로 구성된 시스템을 의미한다. 이러한 소재 시스템의 개념은 많은 경우 소재의 성능이 소재 시스템으로 만들어져 평가된다는 연구개발의 특성을 반영하고 있다. 소재 시스템의 개념을 소재 데이터 구조에 도입한다는 것은 데이터 구조 내에 개별 소재의 정보와 더불어 소재 시스템의 정보를 포함할 수 있다는 의미를 갖는다. 많은 경우, 소재의 물성은 주변 소재와의 상호작용에 의

7) https://www.rd-alliance.org/groups/working-group-international-materials-resource-registries.html
8) 오창석, 김수현, 박지원, 이호원, 임창동, 미래소재 연구데이터 플랫폼 구축사업 상세기획 연구 보고서, 2019.4.15. 과학기술정보통신부

해 영향을 받는 것이 잘 알려져 있다. 대표적인 예로는 박막 소재의 물성이 모재와의 계면 구성에 의해 크게 달라진다는 점을 들 수 있다. 표 1은 본 연구에서 정의하는 소재 시스템의 몇 가지 예를 보여주고 있다.

촉매소재 연구에서는 촉매 입자가 촉매 담지체, 그리고 필요에 따라 촉진제 등의 소재들과 함께 촉매 시스템을 구성하여 촉매 소재로서의 성능을 평가하게 된다. 이 경우 소재 데이터는 촉매 입자를 포함한 구성 소재 각각의 데이터와 구성 소재들로 구축되는 촉매 시스템의 데이터로 구성된다. (소재 데이터의 구성에 대해서는 앞으로 좀 더 상세히 설명될 것이다.) 또 다른 예인 복합소재는 강화 화이버 소재와 매트릭스 소재로 구성되며 이들이 구조적으로 결합되어 복합재의 성능이 평가된다. 센서 소재는 센싱 소재와 센서 시스템을 구성하는 전극과 모재, 그리고 보호막 소재 등으로 구성되어 센서 소재의 성능이 평가된다.[9] 이들 예에서 보는 바와 같이 소재의 성능이 평가되는 단위 시스템을 소재 시스템으로 정의하고, 여기에 사용되는 소재의 데이터와 소재 시스템의 데이터로 구조화하였다. 그러나, 이러한 데이터 구조가 복수의 소재로 구성된 소재 시스템의 데이터만을 수용하는 것은 아니다. 소재 시스템의 정보 없이 단일 소재의 데이터로 구성되면 개별적인 소재 데이터도 수용할 수 있도록 되어 있다. 예를 들어 표 1의 특수강의 경우는 하나의 구성 소재 데이터만을 가지며 소재 시스템의 정보가 존재하지 않는다.

소재 시스템의 개념을 도입한 데이터 구조는 많은 장점을 가지고 있다. 특정 응용 분야에서 소재의 성능은 해당 응용 시스템을 구성하는 다른 물질들과의 물리 화학적 구성 혹은 상호 작용과 연관되어 있다. 본 데이터 구조에서는 소재의 성능이 소재 시스템의 구성에 따라 변화하는 시스템 의존성을 담을 수 있다. 또한, 광범위한 소재의 대부분을 수용할 수 있는 유연하고 포괄적인 구조라고 할 수 있다. 앞서 설명한 특수강의 경우처럼 단일 소재만으로 소재 데이터가 구성되는 경우에도 해당 소재의 데이터를 통일된 데이터 구조 속에 수용할 수 있다. 반대로 구성 소재의 개수에 제한이 없는 소재 시스템의 데이터 구조도 구축할 수 있다. 다만, 소재군별로 소재 시스템의 구성 소재를 정의하는 과정은 해당 소재군 연구에 대한 전문적인 지식이 반영되어야 한다. 따라서, 국가소재연구데이터센터에서 분야별로 소재 시스템의 소재 구성과 소재 시스템을 정의하는 템플레이트를 구축하여 공개할 예정이다. 향후 이러한 템플레이트를 바탕으로 소재 데이터를 수집, 관리, 활용할 다양한 플랫폼이 구축될 수 있을 것이다.

표 1. 본 연구에서 도입된 소재 시스템 개념의 예

소재시스템	구성소재 1	구성소재 2	구성소재 3	구성소재 4	...
촉매소재	촉매입자	담지체	촉진물질		
복합소재	강화화이버	매트릭스소재			
센서소재	센싱소재	전극재	기판	보호막	
memristor 소재	활성층	상층	하층	기판	전극소재
이차전지 양극재	양극재1	양극재2	전극재	전해질	
크롬도금강	모재강	크롬코팅층			
특수강	특수강소재				

9) 이 예시들은 실제 데이터의 구성과 다를 수 있다. 실제 데이터의 구성은 국가소재연구데이터센터 산하 소재군별 전문가로 구성된 특화센터에서 별도로 정의하여 데이터를 수집하고 관리할 것이다.

그림 1은 소재 시스템의 개념을 채택하는 소재 데이터의 일반적인 구조를 보여주고 있다. 일반적으로 소재 데이터는 metadata와 materials, 그리고 system의 3개 영역으로 구성된다. 첫 번째 영역은 소재 데이터의 메타데이터 (metadata) 영역이다. 이 영역에서는 해당 데이터의 명칭 (data name), 데이터 작성자 정보 (contributor), 데이터 작성 연월일 (data generation date), 그리고 데이터에 대한 특이사항 (note on data) 등을 제공한다.

두 번째 영역은 소재의 개별적인 데이터 영역이다. 이 영역은 각 소재 별로 materials_n 으로 표시하는 데이터 그룹을 둔다. 여기서 n은 n번째 소재 데이터 그룹의 인덱스 번호로 정수를 부여한다. 소재 시스템을 구성하는 소재의 갯수만큼 소재 데이터 그룹을 가질 수 있다. 표 1 의 촉매 시스템을 예로 들면, materials_1 데이터 그룹는 촉매 입자에 대한 데이터, materials_2 데이터 그룹은 담지체에 대한 데이터, 그리고 materials_3 데이터 그룹은 촉진물질에 대한 데이터를 갖는다. 각 소재 데이터 그룹은 통상적으로 사용되는 소재의 이름 (name)과 화학적 조성과 구조에 관한 화학적 정보 (chemical information), 소재의 합성 공정 정보 (process), 이론적 연구의 데이터라면 이를 위한 구조 모델 (model), 그리고 소재의 물리 화학적 특성 정보 (property) 로 구성된다.

세 번째 영역인 소재 시스템 (system) 은 해당 소재 시스템에 대한 데이터들을 갖는다. 시스템의 설명 (description) 데이터에 이어, materials 영역에서 정의된 소재들로 이루어진 시스템의 구성 (configuration) 데이터가 제공된다. 이 'configuration' 데이터를 통해 시스템에서 차지하는 각 소재의 역할이 정의된다. 이어서 시스템의 제조 공정 데이터 (process)와 소재 시스템의 성능 데이터 (performance) 로 시스템의 데이터가 구성된다.

그림 2는 그림 1의 데이터 구조를 반영한 소재 데이터의 구성을 좀 더 구체적으로 보여주는 그림이다. 각 소재의 공정정보는 각각의 소재에 대해 일련의 단위 공

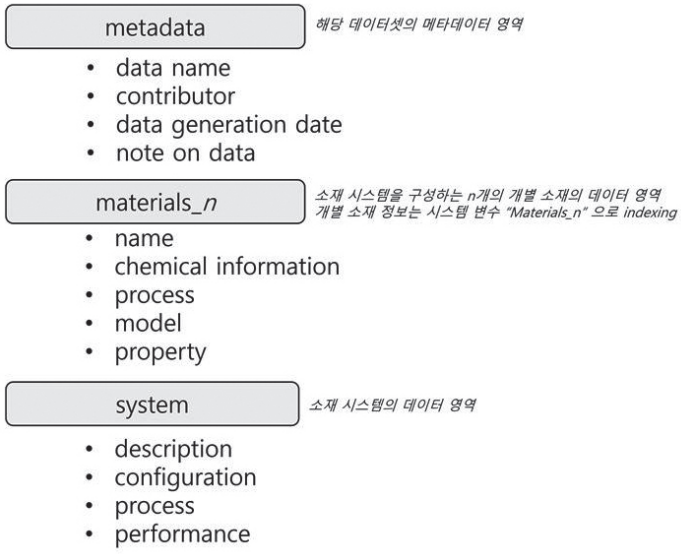

그림 1. 소재 데이터의 일반적인 구조

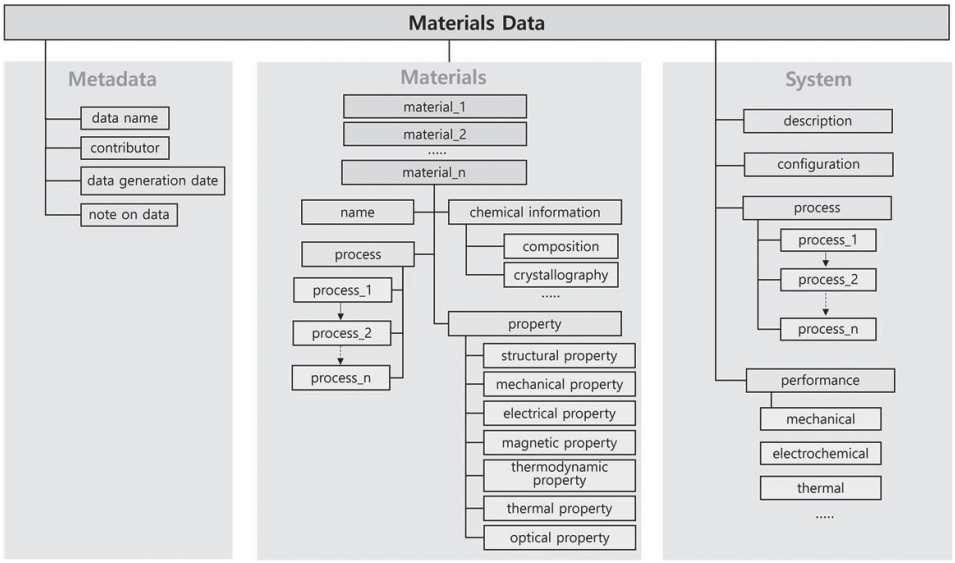

그림 2. 소재 데이터의 구성

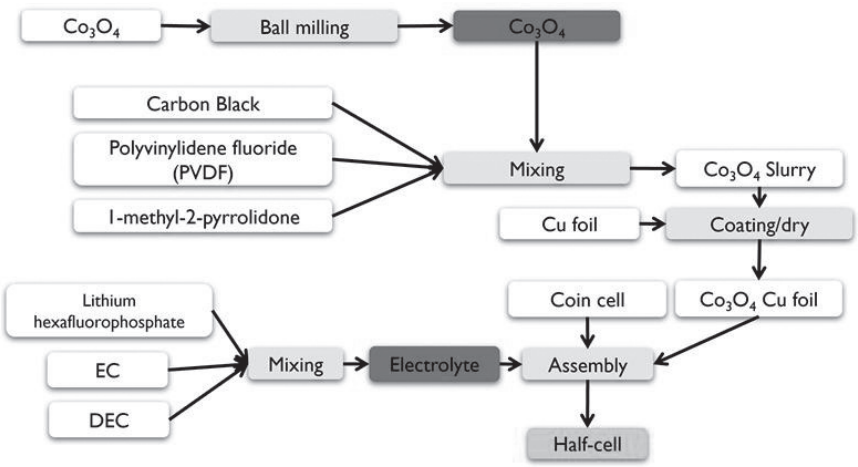

그림 3. Co_3O_4 이차전지 음극 연구를 위한 소재 공정도의 예

정들로 표현할 수 있다고 가정하고, 단위 공정의 순서와 공정조건을 함께 저장하도록 하였다. 이는 소재 시스템에도 마찬가지로 적용하여 소재 시스템을 구축하는 공정 정보를 제공하도록 하였다. 본 연구에서 채택한 소재 시스템의 개념은 복잡한 공정 데이터를 다룰 때 매우 유용하였다. 복잡해 보이는 전체 공정도 개별 소재와 소재 시스템 별로는 일련의 단위 공정들로 표현할 수 있기 때문이다. 그림 3은 Li 2차 전지의 anode로서 Co_3O_4 소재를 연구하는 경우의 소재 공정도이다. 이 예를 가지고 본 데이터 구조에서 공정의 표현

을 만들면 다음과 같다. Anode 소재의 성능 평가를 배터리 half-cell을 만들어 수행한다면, 여기서는 half-cell이 소재 시스템이 될 수 있다. 이 소재 시스템을 구성하는 구성 소재는 소재 시스템의 'configuration'에서 정의되는 소재들로 구성한다. 한 예로 시스템 'configuration'에서 음극활물질과 전해질 만을 정의한다면, materials_1 은 Co_3O_4 음극활물질, materials_2는 electrolyte로 정의된다. 이때 materials_1의 공정은 Co_3O_4 분말의 ball milling이며, materials_2의 공정은 Li hexafluorophosphate, EC 그리고 DEC의 mixing의 단일 공정으로 구성된다. 한편, 소재 시스템의 공정 정보는 ① Co_3O_4 분말 (materials_1)과 전도체 및 바인더를 섞어 Co_3O_4 slurry를 만들기 위한 mixing 공정, ② Cu foil 에의 코팅과 건조, ③ coin cell을 이용하여 electrolyte (materials_2) 와 half-cell을 조립하는 세 개의 단위 공정으로 구성할 수 있다.

소재 시스템의 'configuration' 정보는 어떤 소재가 시스템 내에서 어떠한 구조를 가지고 어떠한 역할을 하는지를 정의한다. 소재 시스템을 구성하는 소재들의 모든 정량정보 (크기, 분율 등) 들도 'configuration'에 포함된다. 따라서, 'configuration' 정보는 시스템 데이터 영역과 소재 데이터 영역의 연결고리 역할을 하며, 소재군에 따라 고유한 키워드 구성을 가진다. 즉, 촉매 시스템의 'configuration'과 배터리 전극 시스템의 'configuration', 복합소재 시스템의 'configuration'들이 서로 다른 응용 분야에서 그 특성을 평가하게 되므로 서로 다른 구성을 하게 된다. 그림 3의 예에서 half cell 시스템의 'configuration' 정보는 anode 활물질과 전해질에 해당하는 소재, 활물질의 함량, anode 전체 질량과 두께, half-cell의 크기와 모양의 정보를 가질 것이다. 한편, 복합구조 소재 시스템의 경우에는 'configuration' 정보가 강화재와 매트릭스로 사용되는 소재를 정의하고 강화재의 함량 등을 포함할 것이다. 소재 시스템의 performance도 시스템마다 평가하는 항목이 다르므로 시스템 별로 특화된다. Materials 데이터 영역의 property는 개별 소재 자체의 고유 물성이어서 소재 시스템에 의존하지 않는 일반적인 특성이지만, system 데이터 영역의 performance는 주어진 system 환경 하에서 특징적으로 발현되는 해당 소재의 성능이라고 할 수 있다.

III. 표준어휘 사전의 구성

본 연구에서 구축하는 소재 데이터 표준어휘 사전은 그림 1의 소재 데이터 체계 하에서 사용되는 키워드들의 정의와 이에 해당하는 데이터의 특성을 담고 있다. 2021년 10월 현재 소재 데이터 표준어휘 사전은 크게 아래 5개의 어휘군과 수치 데이터 표현으로 구성되어 있다.

- 메타데이터 표준어휘군 (metadata vocabulary)
- 소재 표준어휘군 (materials common vocabulary)
- 소재 시스템 표준어휘군 (system specific vocabulary)
- 소재 분석 공통어휘군 (common vocabulary for analysis method)
- 소재 공정 공통어휘군 (common vocabulary for materials process)
- 수치 데이터 표현 (numeric data expression)

여기서 메타데이터 표준어휘군, 소재 표준어휘군, 소재 시스템 표준어휘군은 각각 앞 절에서 기술한 소재 데이터 영역에 해당되는 어휘들을 정의하고 있다. 한편, 소재 분석 공통어휘군, 소재 공정 공통어휘군은 각각 분석 조건 정보와 단위 공정 정보의 어휘들을 정의하고 있다. 이 어휘들은 모든 소재나 소재 시스템에 공통적으로 사용될 수 있어서 별도의 공통어휘군으로 정리하였다. 실제 소재 데이터를 구성할 때에는 필요한

부분에 공통어휘군의 해당 키워드를 채택하여 사용하는데 이는 소재 데이터 템플레이트에서 반영한다.

소재 데이터 내의 모든 측정 및 계산 데이터는 원칙적으로 그 데이터를 만든 측정 혹은 계산 조건을 담은 'measurement' 데이터와 함께 저장된다. 이때 사용되는 분석/계산 방법 데이터에 대한 모든 표준어휘는 본 사전의 소재 분석 공통어휘군에 방법 별로 정의되어 있다. 따라서, 소재 데이터 내의 'measurement' 영역은 '소재 분석 공통어휘군'에 정의된 해당 항목을 삽입하여 구성한다. 사용한 측정 혹은 계산 방법이 다수이면 복수 개의 분석 방법 데이터들을 임의의 순서대로 삽입한다. 그림 4의 데이터 예에서 소재의 결정구조 정보 중 lattice parameter 데이터는 XRD와 TEM을 통해 측정되었고, 구체적인 측정 조건은 XRD 와 TEM object 내에 저장되어 있음을 보여 준다.

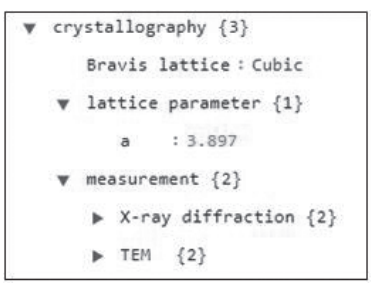

그림 4. 소재 데이터의 결정구조 정보 중 lattice parameter 데이터의 예

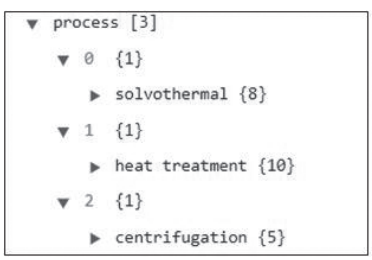

그림 5. 순차적인 3개의 단위 공정으로 이루어진 소재 공정 데이터의 예

그림 2에서 보는 바와 같이 소재 연구 데이터 체계의 'process' 데이터 영역은 단위 'process'의 sequence로 구성된다. 'process' 데이터 영역의 공정 정보는 실제 공정의 순서에 맞춰 array로 저장된다. 그림 5는 solvothermal 'process'로 합성한 뒤 heat treatment를 거쳐 centrifugation에 의해 만들어지는 소재의 공정 정보의 예시이다. 각 공정 조건 데이터에 사용되는 키워드 표준어휘는 "소재 공정 공통어휘군"에 정의되어 있다.

수치 데이터 표현 (numerical data expression) 에서는 단일 수치 데이터인 경우와 x-y 테이블 데이터인 경우 표현방식들을 정의하고 있다. 표현 방식에 관계없이 측정이나 분석에 의해 생성된 수치 데이터는 모두 측정값 (value)과 측정 분산 수준을 나타내는 불확도 (uncertainty)를 갖는다. 여기서 'uncertainty'의 단위는 수치 데이터 'value'의 단위와 동일하다. 그림 6은 단일 수치 데이터의 예로서 temperature의 'value'와 'uncertainty'를 보여주고 있다.

한편, 표 2의 예와 같이 테이블로 주어지는 수치 데이터는 independent parameter x 값의 의미와 단위를 'x_definition' object에서 정의하고, 해당 수치 데이

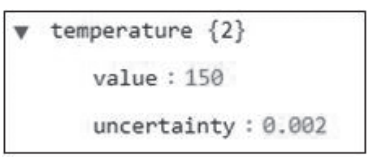

그림 6. 단일 수치 데이터 temperature의 예 (temperature의 단위는 어휘사전에서 정의한다.)

표 2. array 수치 데이터의 예

x	x_uncertainty	value	value_uncertainty
130	0.5	0.8	0.001
180	0.5	1.3	0.001
240	0.5	2.2	0.001

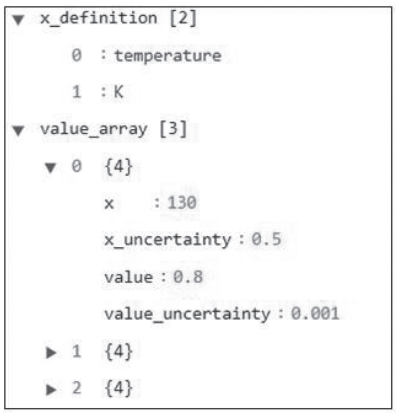

그림 7. 표 2의 수치 데이터 array의 데이터 구조

터를 테이블의 한 줄씩 'value_array'에 순차적으로 저장하도록 하였다. (그림 7 참조)

표준어휘 사전에서 각 어휘는 다음과 같은 5개의 항목을 가지고 있다.

1. eng_definition : 키워드 어휘의 영문 설명
2. var_name : 소재 데이터베이스 내에서 사용되는 어휘의 변수명
3. alias : 키워드 어휘와 동일한 의미로 사용되는 용어
4. data_type : 해당 데이터의 형태 (예: 수치 데이터, 텍스트 데이터 등)
5. data_unit : 데이터 타입이 수치 데이터인 경우 데이터의 단위
6. data_example : 해당 데이터의 예시

그림 8 은 사전 내 어휘의 한 예로서 materials 데이터 영역의 property 데이터 중 thermal property 부분을 보여주고 있다. 이 경우 어휘 사전에는 thermal property의 영문 정의 (eng_definition) 와 변수명 (var_name), 동의어 (alias)가 정의되어 있다. thermal property는 데이터의 상부 구조로서 그 자체로는 데이터를 갖지 않으므로 data_type, data_

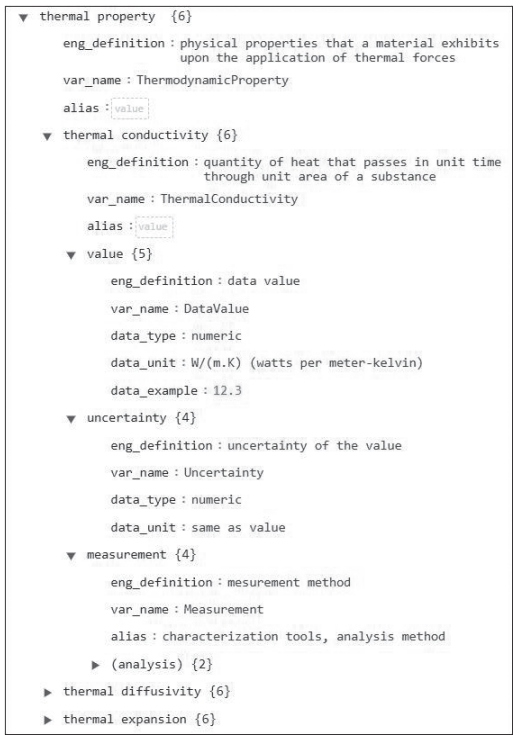

그림 8. 소재 연구 데이터 표준어휘 사전에 기재되어 있는 어휘의 예: materials 데이터 영역의 property 데이터 중 thermal property 관련 어휘들이 정의되어 있는 부분

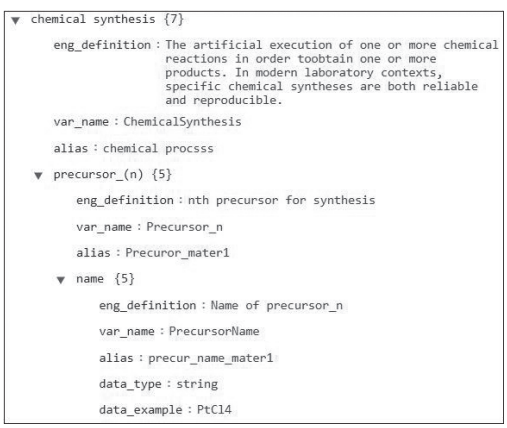

그림 9. 소재 연구 데이터 표준어휘 사전에 기재되어 있는 어휘의 예: chemical synthesis 소재공정 데이터의 어휘들이 정의되어 있는 부분

표 3. 소재데이터 표준어휘 사전에서 정의하는 키워드 표준어휘 목록

어휘군	category 1	category 2	category 3
metadata vocabulary	data name		
	contributor	name	
		affiliation	
	data generation date		
	note on data		
materials common vocabulary	materials_(n)	name	
		chemical information	composition
			crystallography
			SMILES
		process	(process_n)
		model	elements
			phase
			box dimension
			periodic boundary condition
			structure file
		property	structural property
			mechanical property
			electrical property
			magnetic property
			thermodynamic property
			thermal property
			optical property
			corrosion property
system specitic vocabulary	system_catalyst	description	
		configuration	active material
			amount of active material
			promotor
			amount of promotor
			support materials
		process	(process_n)
		performance	electrochemical
			thermal
	system_porous_materials	description	
		configuration	framework_(n)
			inserted molecule
		process	(process_n)
		performance	gas adsorption

어휘군	category 1	category 2	category 3
system specitic vocabulary	system_memristive	description	
		configuration	electrode_1
			buffer_1
			active layer_(n)
			buffer_2
			electrode_2
		process	(process_n)
		performance	current
			endurance
			mechanism
			operating speed
			resistance
			retention
			selectivity
			voltage
common vocabulary for analysis method	creep test		
	DFT		
	electrochemical activity		
	empirical MD		
	fatigue test		
	gas adsorption/desorption isotherm		
	gas chromatography		
	hardness test		
	impact test		
	infrared spectroscopy		
	memristive activity		
	nuclear magnetic resonance		
	optical microscopy		
	TEM		
	tensile test		
	thermal activity		
	thermogravimetric		
	x-ray diffraction		
common vocabulary for materials process	centrifugation		
	chemical mechanical polishing		
	chemical synthesis		
	Czochralski growth method		
	drying		
	electrochemical deposition		

어휘군	category 1	category 2	category 3
common vocabulary for materials process	epitaxy		
	exfoliation		
	thermomechanical process		
	hydrothermal growth method		
	ion intercalation		
	Kyropoulos growth method		
	lithography_electron-beam		
	lithography_photon		
	microwave-assited method		
	molecular beam epitaxy		
	polishing		
	pulsed laser deposition		
	rinsing		
	self-assembly		
	sintering		
	skull crucible growth method		
	sol-gel synthesis		
	sonochemical synthesis		
	solvothermal process		
	sonication		
	sputter deposition		
	Verneuil growth method		
	wet etching		
numeric data expression	single numeric data		
	array numeric data		

unit, data_example은 정의하지 않는다. 실제 데이터는 thermal property의 하부 데이터인 thermal conductivity, thermal diffusivity, thermal expansion 등에 존재한다. 그림 8의 예에서 thermal conductivity 는 단일 수치 데이터로서 "6. 수치데이터 표현"에 정의된 value와 uncertainty 키워드를 갖는다. 실제 수치 데이터는 value와 uncertainty로 구성되므로 이들 키워드에는 해당 데이터의 형태와 단위가 정의되어 있다. 측정된 특성 데이터에는 특성의 측정 정보가 measurement 영역에 분석 방법 별로 제공된다. Measurement 영역의 (analysis) 항은 "소재 분석 공통어휘군"에서 해당 분석법의 이름으로 대치되며, 해당 분석법의 표준 어휘들을 적용하여 작성한다. 그림 9는 합성을 위한 화학공정 (chemical syntehsis)의 데이터 키워드 중 precursor 부분을 보여주고 있다. 여기서 precursor_(n)의 name은 문자열 (string) 데이터를 가지므로 이 어휘에는 data type과 data example만 정의되어 있다.

2021년 11월 현재 공개된 표준어휘 사전은 json 파일과 일반 문서의 형태로 공개되어 있으며[10] 본 논문의 보충 자료로 첨부하였다. 현재 어휘 사전에서 정의하고 있는 키워드 어휘를 표 3에 정리하였다. 표 3에는 데이터 구조의 category 3까지 정리하였으며, 키워드에 대한 세부적인 정보는 본 고의 보충자료인 2011.11월자 소재표준어휘에 기재되어 있다.

III. 결언 및 제안

소재 연구 데이터의 체계적인 수집, 관리, 활용을 위한 표준 데이터 구조와 키워드의 표준어휘를 구축하여 공개하였다. 광범위한 영역의 소재 데이터들을 일관된 개념의 데이터 구조로 수용하기 위해 "소재 시스템"의 개념을 도입하여 데이터를 구조화시켰다. 그 결과, 소재 연구 데이터는 메타데이터, 소재 데이터 그리고 시스템 데이터들로 구성하였다. 이러한 데이터 체계를 바탕으로 통일된 키워드 어휘를 정의하고 각 키워드에 해당하는 데이터의 특성을 정의한 표준어휘 사전을 발간하였다. 그러나, 본 고에서 논한 소재 연구 데이터의 표준 키워드 어휘 사전은 아직 모든 소재군에 대한 키워드를 담은 사전으로 완성된 것은 아니다. 앞으로도 지속적으로 소재의 영역을 확장하고 이에 필요한 키워드들을 정의하는 보완작업이 진행되면서 모든 소재데이터를 아우르는 표준어휘 사전으로 발전할 것이다.

10) 논문의 작성 시점에 열람 가능한 사전은 2021년 8월 23일에 작성되어 소재연구데이터 표준어휘 전문위와 소재어휘 표준화 위원회의 승인을 받은 사전이다. 표준어휘사전은 지속적으로 갱신되고 있으며, 최근 사전의 정보는 표준어휘사전관리플랫폼 (https://matdict.rnddata.org) 에서 제공하고 있다.